AN EVALUATION OF THE USEFULNESS OF THE SMALL HEAT SHOCK GENES FOR PHYLOGENETIC ANALYSIS IN PLANTS[1]

Elizabeth R. Waters[2]

ABSTRACT

Sequences from chloroplast DNA and rDNA genes have proved useful in estimating phylogenetic relationships among plants. However, the coding regions of rDNA evolve slowly and are not useful at lower levels of analysis. There are chloroplast-encoded genes that evolve at a higher rate, but since chloroplast DNA is uniparentally inherited, nuclear markers are also needed when introgression or hybridization have occurred. There is then a need for additional nuclear genes that can be used in phylogenetic analysis. Nuclear genes have been largely unused in phylogenetic analysis for several reasons, including the difficulty of distinguishing orthologous from paralogous genes and the problems that can ensue from constructing phylogenetic trees from data matrices that contain both types of genes. This paper is an evaluation of the potential of the genes encoding the small heat shock proteins for phylogenetic analysis of plants. The small heat shock genes in plants are a super gene family composed of four gene families. In this paper I present restriction site analysis of the small heat shock genes from the Brassicaceae, and similarity, rate of evolution, and phylogenetic analysis of small heat shock gene sequences from monocots and dicots. I show that these genes possess the necessary genetic variation for phylogenetic analysis and that the gene families are easily distinguishable from each other, and that, in at least some of the gene families, orthologous and paralogous genes do not pose a problem for phylogenetic analysis.

In recent years the availability of molecular data has vastly increased our ability to test hypotheses of both phylogenetic relationships among plant groups and population-level evolutionary processes. These studies, with few exceptions, have relied primarily on chloroplast DNA (cpDNA) and nuclear ribosomal DNA (rDNA) (Clegg & Zurawski, 1992; Hamby & Zimmer, 1992). The large numbers of cpDNA sequences (Chase et al., 1993) across many taxonomic groups have enabled scientists to test hypotheses concerning rates of evolution of genes used as markers in phylogenetic analysis, in addition to addressing phylogenetic questions (Bousquet et al., 1992; Gaut et al., 1992). However, the recognition of the distinction between gene trees and species trees and the effects that hybridization can have on phylogenies constructed from uniparentally inherited genes (Dorado et al., 1992; Doyle, 1992; Pamilo & Nei, 1988; Rieseberg et al., 1990; Rieseberg & Brunsfeld, 1992; Rieseberg & Soltis, 1991) has led to the awareness that additional molecular markers are needed. In addition, there is concern over the effects of multiple changes at one site (Smith, 1989) and secondary structure and compensatory mutations (Dixon & Hillis, 1993) on phylogenetic reconstruction using rDNA sequences.

Systematists have been wary of using nuclear genes for phylogenetic analysis because many nuclear genes are members of multi-gene families (Doyle, 1993). The inability to distinguish orthologous and paralogous genes within a multi-gene family could undermine attempts to construct organismal relationships. Gene conversion and recombination within gene families will obscure organismal relationships, and this may not be immediately obvious from the cladograms (Sanderson & Doyle, 1992). Thus, prior to its use in phylogenetic analysis the evolutionary dynamics of a gene should be examined. In this paper the small

[1] This work was partially supported by a Sigma Xi Grant-in-Aid of Research from the Department of Biology at Washington University, and the Research Training Program for the Analysis of Biological Diversification at the University of Arizona. B. Schaal at Washington University and Elizabeth Vierling at The University of Arizona provided supply money, laboratory equipment, and stimulating intellectual environments. J. Bergelsen, A. Colwell, T. D. Ho, R. Kranz, B. Schaal, and A. Templeton provided useful suggestions on an earlier version of this manuscript. This paper was greatly improved by the comments and suggestions of Elizabeth Kellogg, Pamela S. Soltis, and an anonymous reviewer.

[2] Department of Biology, Washington University, St. Louis, Missouri 63130, U.S.A. Address for correspondence: Department of Biochemistry, University of Arizona, Tucson, Arizona 85721, U.S.A.

heat shock protein genes are examined for their usefulness in phylogenetic reconstruction.

Organisms respond to heat stress with the rapid induction of the heat shock genes and subsequent production of heat shock proteins (hsp) that are necessary for thermal tolerance. The heat shock genes are members of multi-gene families based on the size of the proteins the genes encode. The small heat shock proteins are less than 30 kDa (kilo-Dalton) in size. The large heat shock proteins range from 60 to 100 kDa. The 90, 70, and 60 kDa proteins are the most extensively studied of all of the heat shock proteins. The large heat shock proteins are chaperone proteins, i.e., they help in the proper folding and translocation of other proteins (Becker & Craig, 1994; Gething & Sambrook, 1992). The function of the small heat shock proteins, however, remains elusive. In most organisms, other than plants, the majority of the heat shock proteins produced during stress are large. The small heat shock proteins are present, but are a small percentage of the total protein. In *Drosophila* there are just four small heat shock genes, and in *Saccharomyces* there is only one (Lindquist & Craig, 1988; Susek & Lindquist, 1989). In contrast, plants produce a variety of small heat shock proteins (Vierling, 1991). There are four small heat shock protein gene families in plants. The four classes are based on overall sequence similarity and cellular localization. There are two classes of proteins that localize to the cytosol, Class I and Class II. The third class of proteins is found in the endoplasmic recticulum, and the fourth in the chloroplast. The carboxyl-terminal domain is more conserved than the amino-terminal domain both within and across protein classes (Vierling, 1991). Vierling (1991) suggested that the four classes of small heat shock genes in plants may have evolved through gene duplications.

Materials and Methods

Cloning and Mapping of a Small Heat Shock Gene in *Brassica nigra*

Brassica nigra (CrGC accession number 2-1) seeds were obtained from the Crucifer Genetic Cooperative (CrGC, Madison, Wisconsin). Seeds were germinated and planted in Terra-Lite (Grace Horticultural Products, Cambridge, Massachusetts) and grown in growth chambers in the Washington University Biology Department Plant Growth Facility (St. Louis, Missouri). Plants were harvested when they were two to three weeks old. Leaf tissue was ground to a powder in liquid nitrogen and stored at −80°C. Total genomic DNA was isolated from individuals using a modified CTAB procedure with a phenol-chloroform extraction (King & Schaal, 1990).

Genomic DNA from *Brassica nigra* was cut with *Xba* I and was ligated into Lambda Zap vector (Stratagene, San Diego, California) also cut with *Xba* I. The lambda library was amplified in NM522 cells obtained from Stratagene. Lambda plaques were transferred from agar plates to nylon filters by the filter-lift procedure (Sambrook et al., 1989). Plaques of the heat shock gene were identified in the library using a cDNA clone of a small heat shock gene from *Glycine max* pCE53 (Schoffl & Key, 1982). The nylon (Magna nylon, MSI) filters were pre-hybridized and hybridized with hexamer-labeled pCE53 using standard protocols (King & Schaal, 1990) at moderate stringency: 2 × 15 min. with 2 × SSC 0.1% SDS at 25°C and 2 × 25 min. with 1 × SSC 0.1% SDS at 37°C (Sambrook et al., 1989). Plaques were isolated and pBluescript plasmids were obtained from the plaques according to the Stratagene in-vivo excision protocol.

Plasmid DNA was obtained by alkaline lysis followed by phenol extraction (Sambrook et al., 1989) and digested with the following restriction enzymes: *Acc* I, *Hind* III, *Sal* I, *Pst* I, *Eco* RI, *Xba* I, and *Xho* I (New England Biolabs). Reactions were conducted according to manufacturer's recommendations. Restriction digests were electrophoresed in 0.85% agarose gels in tris-acetate buffer. The DNA was visualized under UV light by staining with ethidium bromide. The DNA was transferred to nylon filters using standard Southern blotting procedures. The coding region of the genes was identified by hybridization of the plasmid DNA digests with radio-labeled cDNA clone of a 17.6 kDa hsp from *Arabidopsis thaliana* (Helm & Vierling, 1989).

Survey of Genetic Variability

Genomic DNA was isolated, using the CTAB extraction procedure described above, from each of 15 individuals of *Brassica nigra* (CrGC #2-1). Genomic DNAs were digested with *Eco* RI, and the restriction digests were electrophoresed in 0.8% agarose gels in tris-acetate buffer and transferred to nylon membrane using standard procedures. Nylon filters were hybridized with the hexamer-labeled small heat shock gene clone from *B. nigra*, described above. Filters were washed 2 × 15 min. with 2 × SSC 0.1% SDS at 65°C and 2 × 45 min. with 0.1 × SSC 0.1% SDS at 65°C and

exposed to x-ray film with intensifying screens at −80°C for three to five days.

To assess within-population restriction site variation in the small heat shock genes, seeds of *Brassica oleracea*, *Cardamine keyserri*, and *Rorippa schlechteri* were collected in Papua New Guinea. Collections of seeds (at least 20 seeds from each of 10 individuals per population) were made in and around the village of Safalitikin, Urapmin, Sandaun Province. Voucher specimens were deposited in the herbarium at UPNG and at MO.

Seeds were germinated and grown as described above. When plants were approximately four weeks old, leaves were harvested and stored at −80°C. Eight individuals from two populations of *Rorippa schlechteri* and eight individuals from one population each of *Brassica oleracea* and *Cardamine keyserri* were surveyed for within-population variation. DNA from individual plants was digested with *Ban* I, *Eco* RI, *Hpa* I, *Hinf* I, *Hind* III, *Rsa* I, *Pst* I, and *Taq* I (New England Biolabs) according to manufacturer's instructions. Restriction digests were electrophoresed in 0.8–1.2% agarose gels in tris-acetate buffer. DNA was transferred to nylon membranes and hybridized with the hexamer-labeled small heat shock clone from *B. nigra*, described above. Filters were washed 2 × 15 min. with 2 × SSC 0.1% SDS at 65°C and 2 × 45 min. with 0.1 × SSC 0.1% SDS at 65°C and exposed to x-ray film with intensifying screens at −80°C for three to five days.

SURVEY OF THE *BRASSICA* TRIANGLE AND OTHER BRASSICACEAE

Restriction sites in the small heat shock genes were surveyed among the following species: *Rorippa schlechteri*, *Raphanus sativus* (CrGC #7-1), *Cardamine keyserri*, *Arabidopsis thaliana* (CrGC #9-3), *Brassica nigra* (CrGC #2-1), *B. oleracea* (CrGC #3-1), *B. rapa* (CrGC #1-1), *B. carinata* (CrGC #6-1), *B. juncea* (CrGC #4-1), and *B. napus* (CrGC #5-1). The *Brassica* triangle is composed of the diploids *B. nigra*, *B. oleraceae*, and *B. rapa* and the amphidiploids *B. carinata*, *B. juncea*, and *B. napus* (Pakrash & Hinata, 1980).

Genomic DNA from at least two individuals from each species was isolated and cut with *Bam* HI, *Eco* RI, *Fok* I, *Hind* III, *Hae* III, *Pst* I, *Rsa* I, *Sal* I, and *Xba* I restriction enzymes. Digested DNAs were separated in 0.8% agarose gels, transferred to nylon membranes, and hybridized with the hexamer-labeled small heat shock clone from *Brassica nigra*, described above. Filters were washed 2 × 15 min. with 2 × SSC 0.1% SDS at 65°C and 2 × 45 min. with 0.1 × SSC 0.1% SDS at 65°C and exposed to x-ray film with intensifying screens at −80°C for three to five days.

A portion of the gene encoding the small heat shock protein that is localized to the chloroplast was amplified from the genomic DNA of *Arabidopsis thaliana*, *Cardamine keyserri*, *Brassica nigra*, *B. oleracea*, *B. napus*, and *B. juncea* using Taq DNA polymerase, the manufacturer's buffer (Boehringer Mannheim), and primers internal to the coding sequence of the *A. thaliana* gene (Osteryoung et al., 1993). The primers were from base pairs 151–169 and 765–782 (using the start of the coding sequence as base pair 1). The DNA was amplified for 35 cycles with the following conditions: 94°C 1 min., 50°C 2 min., and 72°C 2 min. Products were visualized after electrophoresis in a 0.9% agarose gel with ethidium bromide.

SEQUENCE DATA ANALYSIS

Gene sequences were obtained from GenBank (see Table 1). Sequences were compared using the Gap program in the Genetic Computer Group (GCG) computer package. Alignments were generated using Pileup in GCG and were refined by hand using Lineup (see Appendix). The DNA sequences coding for the transit peptides were not used in these analyses. Synonymous and nonsynonymous substitutions were calculated using the computer program of Li (1993). Phylogenetic analyses of the small heat shock gene sequences were done using PAUP (3.1.1) (Swofford, 1993). Heuristic searches were conducted using 100 random addition taxon replicates with MULPARS and TBR branch swapping. A bootstrap analysis with 100 replicates was performed to assess support for branches.

RESULTS

A full-length genomic clone encoding a cytosolic small heat shock protein was isolated. A restriction map of that clone is presented in Figure 1. The restriction enzyme *Eco* RI cleaves this clone twice. An *Eco* RI digest of genomic DNA of *Brassica nigra* probed with the clone of the small heat shock gene is presented in Figure 2. The sizes of the genomic *Eco* RI fragments in Figure 2 correspond to those predicted from the plasmid restriction map in Figure 1. The results of this Southern blot and others (data not shown) indicate that this gene is present in a single copy in *B. nigra*.

POPULATION-LEVEL VARIATION

To determine if there is variability in the small heat shock gene within populations, I conducted a survey in *Brassica nigra*, *B. oleracea*, *Cardamine*

Genomic heat shock clone from *Brassica nigra*

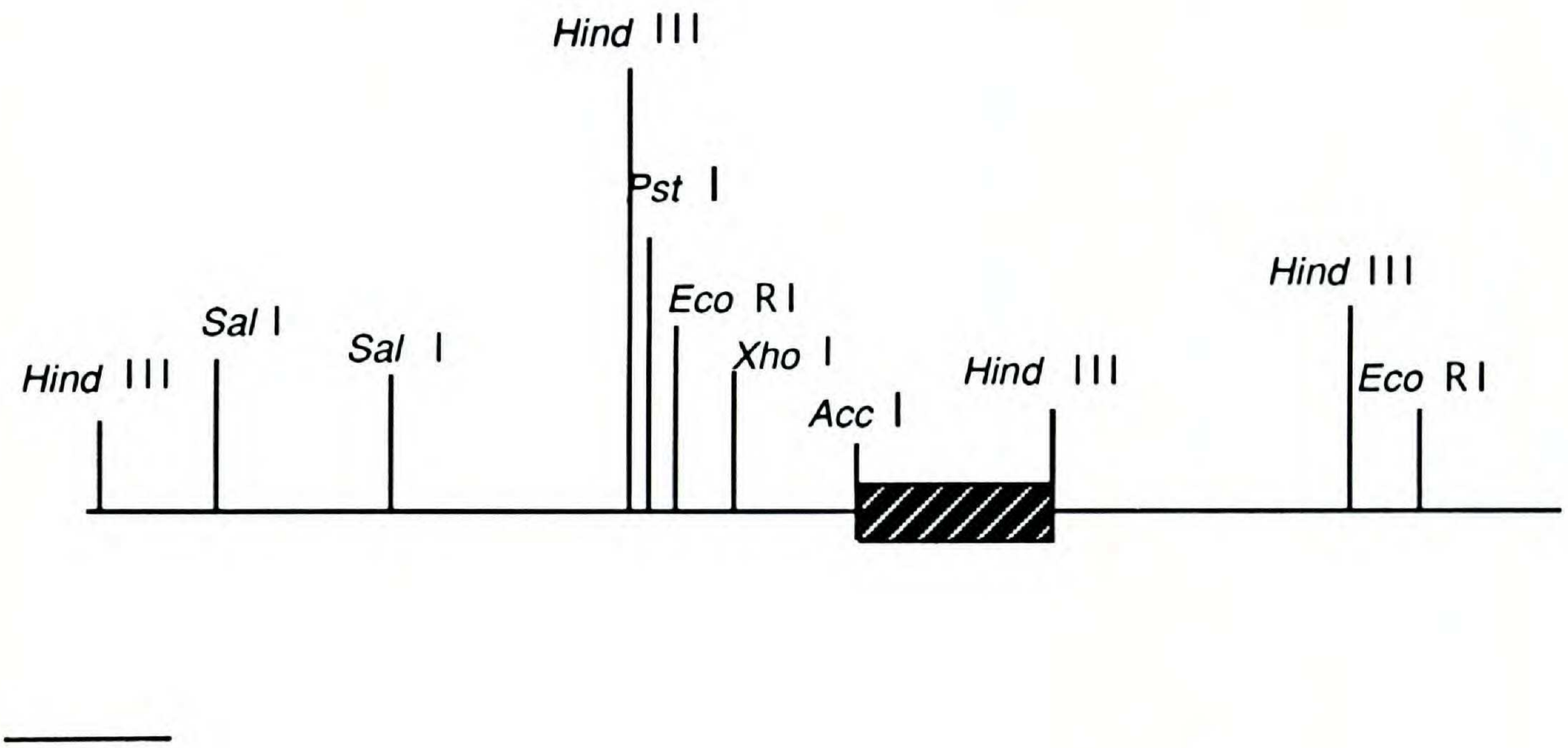

FIGURE 1. Restriction map of a *Brassica nigra* small heat shock gene. The shaded box represents the coding region.

TABLE 1. Sources of sequences used in phylogenetic analyses of small heat shock genes.

Species	Protein	GenBank accession	Reference
A. Genes for the chloroplast-localized proteins:			
Triticum aestivum Percival	Hsp 26A	X58280	Weng et al., 1991
Triticum aestivum	Hsp 26B	X67328	unpublished
Pisum sativum Poir.	Hsp 21	X07187	Vierling et al., 1988
Glycine max Merr.	Hsp 22	X07188	Vierling et al., 1988
Petunia hybrida Vilm.	Hsp 21	X54103	Chen & Vierling, 1991
Arabidopsis thaliana B. Heyne	Hsp 21	X54102	Chen & Vierling, 1991
B. Genes for endomembrane-localized proteins:			
Pisum sativum	Hsp 22	M33898	Helm et al., 1993
Glycine max	Hsp 22	X63198	Helm et al., 1993
C. Genes for the class I cytosolically-localized proteins:			
Zea mays Doebley & Iltis	Hsp 17.2	X65725	unpublished
Pisum sativum	Hsp 18.1	M33899	Lauzon et al., 1990
Arabidopsis thaliana	Hsp 17.6	X16076	Goping et al., 1991
D. Genes for the class II cytosolically-localized proteins:			
Zea mays	Hsp 17.5	X54076	Goping et al., 1991
Zea mays	Hsp 17.8	X54075	Goping et al., 1991
Triticum aestivum	Hsp 17.3	X58279	Weng et al., 1991
Pisum sativum	Hsp 17.7	M33901	Lauzon et al., 1990
Glycine max	Hsp 17.9	X07159	Raschke et al., 1988
Arabidopsis thaliana	Hsp 17.6	X63443	Barling et al., 1992
Pharbatis nil Roth	Hsp 18.8	M99430	Krishna et al., 1992
Pharbatis nil	Hsp 17.2	M99429	Krishna et al., 1992

EcoR I Digest

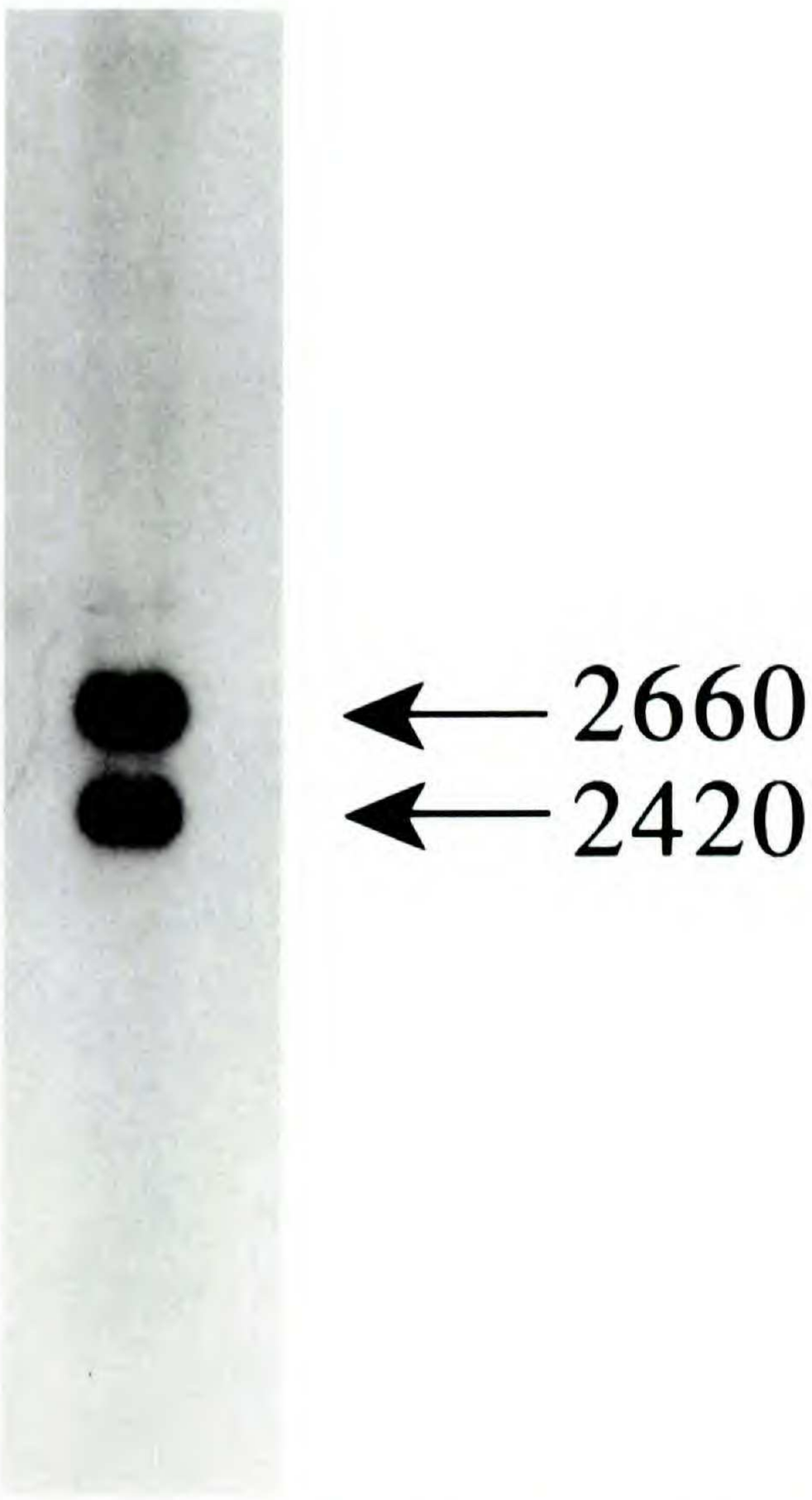

FIGURE 2. *Eco* RI digest of *Brassica nigra* genomic DNA probed with the clone of the *B. nigra* small heat shock gene.

keyserri, and *Rorippa schlechteri*. DNA from 15 individuals of the CrGC rapidly cycling population of *B. nigra* was digested with *Eco* RI. This blot was probed with the full-length heat shock clone (containing both the coding and the flanking regions), and no polymorphisms were detected (data not shown). Populations of *B. oleracea*, *R. schlechteri*, and *C. keyserri* were also surveyed. DNAs from eight individuals from each of these populations were digested with eight restriction enzymes, and the blots were probed with the full-length probe. Again, there was no detectable variation for the small heat shock gene in these populations (data not shown).

VARIATION AMONG MEMBERS OF THE BRASSICACEAE

Restriction site variation in the small heat shock gene was assessed among the members of the *Brassica* triangle and among other genera of the Brassicaceae, *Arabidopsis thaliana*, *Cardamine keyserri*, and *Rorippa schlechteri*. The relationships between the members of the *Brassica* triangle were established with cytological, morphological, and genetic evidence (Pakrash & Hinata, 1980), and analysis of the chloroplast genomes established the maternal and paternal parents of the amphidiploids (Erickson et al., 1983; Palmer et al., 1983). *Brassica carinata* has the chloroplast genome of *B. nigra*, and *B. juncea* has the chloroplast genome of *B. rapa*. Although the data for *B. napus* are equivocal, it appears that in at least some accessions *B. rapa* is the maternal parent. DNAs from each of these species were digested with five restriction enzymes and probed with the full-length clone and with a clone containing only the coding region. No restriction site polymorphisms were found within the coding regions among members of the *Brassica* triangle studied, including *Raphanus* (data not shown). However, variation in restriction sites was detected within the coding region among genera in the Brassicaceae. When the Southern blots were hybridized with the full-length clone, restriction site variation was observed among the members of the *Brassica* triangle, indicating sequence divergence in the flanking region (Fig. 3).

Primers internal to the coding region were used to amplify the gene encoding the small heat shock protein localized to the chloroplast from six species of the Brassicaceae. A single band was present in *Brassica nigra*, *B. oleraceae*, *Cardamine keyserri*, *Arabidopsis thaliana*, *B. juncea*, and *B. napus* (Fig. 4). This indicates that this is a single-copy gene and that PCR (Polymerase Chain Reaction) amplification is easily accomplished.

SEQUENCE ANALYSIS

Sequences from different gene families were compared within *Arabidopsis thaliana* and within *Pisum sativum* (Table 2). In this analysis, the gene for the chloroplast-localized protein in *A. thaliana* was compared to other genes from *A. thaliana* and the gene for the chloroplast protein in *P. sativum* was compared to other genes in *P. sati-*

vum; no interspecific comparisons were made. The gene sequences for the endomembrane proteins and both classes of the cytosolically localized protein are less than 50% similar to the gene for the protein localized to the chloroplast. In contrast, the genes for the chloroplast proteins are more than 50% similar to each other based on interspecific comparisons (Table 3), even between monocots and dicots. Genes from each class or family are more similar to each other than they are to other small heat shock proteins in the same species. It is clear from these comparisons that there is no, or very little, gene conversion across gene families.

Analysis of synonymous (Ks) and nonsynonymous (Ka) substitutions (Table 4) indicates that the small heat shock genes are evolving at rates comparable to that reported for other nuclear genes (Wolfe et al., 1989). In the comparison among the grasses, Ks is not saturated, and Ka is, as expected, much lower than Ks. In the comparisons among the dicots (Table 4), Ks is saturated, reflecting the more distant relationships of these species. The Ka calculated for the small heat shock genes is comparable to the rates reported for the gene for chalcone synthase (Wolfe et al., 1989). Rates of synonymous and nonsynonymous substitutions were determined for three of the gene classes for *Arabidopsis thaliana* and *Glycine max*. The complete sequence of the gene for the chloroplast-localized protein in *G. max* has not been isolated, precluding a comparison of the rate for the entire gene. The gene classes are all evolving at approximately the same rate (Table 4).

In a phylogenetic analysis of the aligned gene sequences of the four major classes of proteins (the Class I and Class II cytosolically localized proteins, the endomembrane proteins, and the chloroplast proteins), the sequences of each gene family form well-supported clades (Fig. 5). Third positions of codons were omitted from the analysis because the rate analysis indicated that substitutions at these positions are saturated. A single tree of 936 steps with a consistency index (CI) of 0.640 was found in all the 100 heuristic searches. The order of gene duplications is not known, so the tree is arbitrarily rooted with the Class I sequences. A tree with the same topology was obtained using aligned amino acid sequences for the same genes (data not shown). Within each gene family, the gene phylogenies are roughly congruent with the species phylogenies. However, the taxon sampling is not the same in each gene family. There is weak support (i.e., bootstrap values less than 50%) for some of the branches within the Class II gene family.

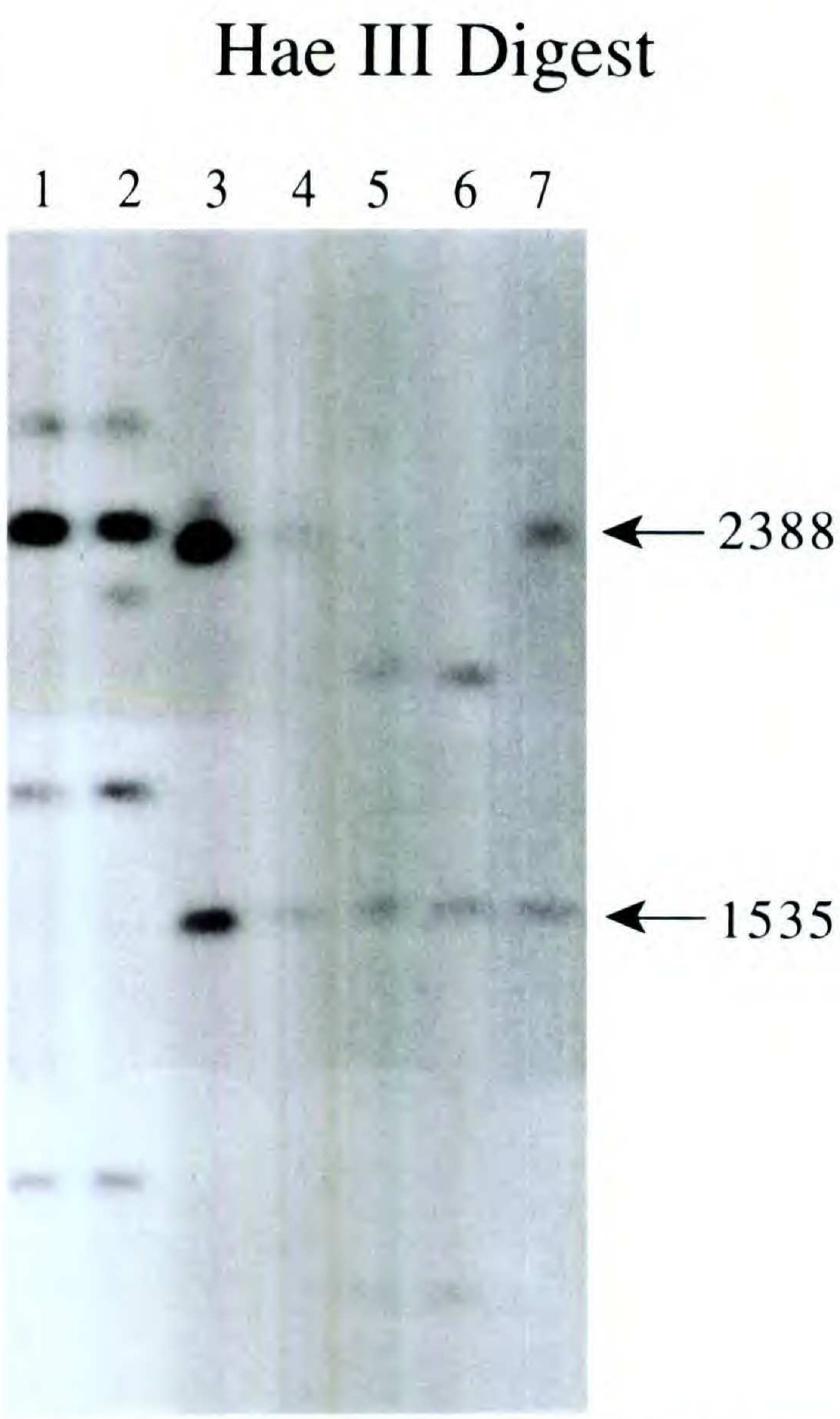

FIGURE 3. *Hae* III digest probed with the *B. nigra* small heat shock gene. Lane 1, *B. nigra*; lane 2, *B. juncea*; lanes 3 and 4, *R. sativus*; lanes 5 and 6, *A. thaliana*; lane 7, *B. oleracea*.

DISCUSSION

Phylogenetic analysis of molecular data sets has enabled evolutionary biologists to approach many questions not answerable with morphological data alone. In plant evolutionary studies, rDNA and cpDNA are the most frequently used molecular markers. Chloroplast DNA is useful in reconstructing organismal relationships above the species level,

TABLE 2. Similarity of the gene coding for the chloroplast-localized protein to other small heat shock proteins in *Arabidopsis thaliana* and *Pisum sativum*.

	A. thaliana	*P. sativum*
Class I	47.9%	44.5%
Class II	45.0%	47.6%
Endomembrane-localized		41.6%

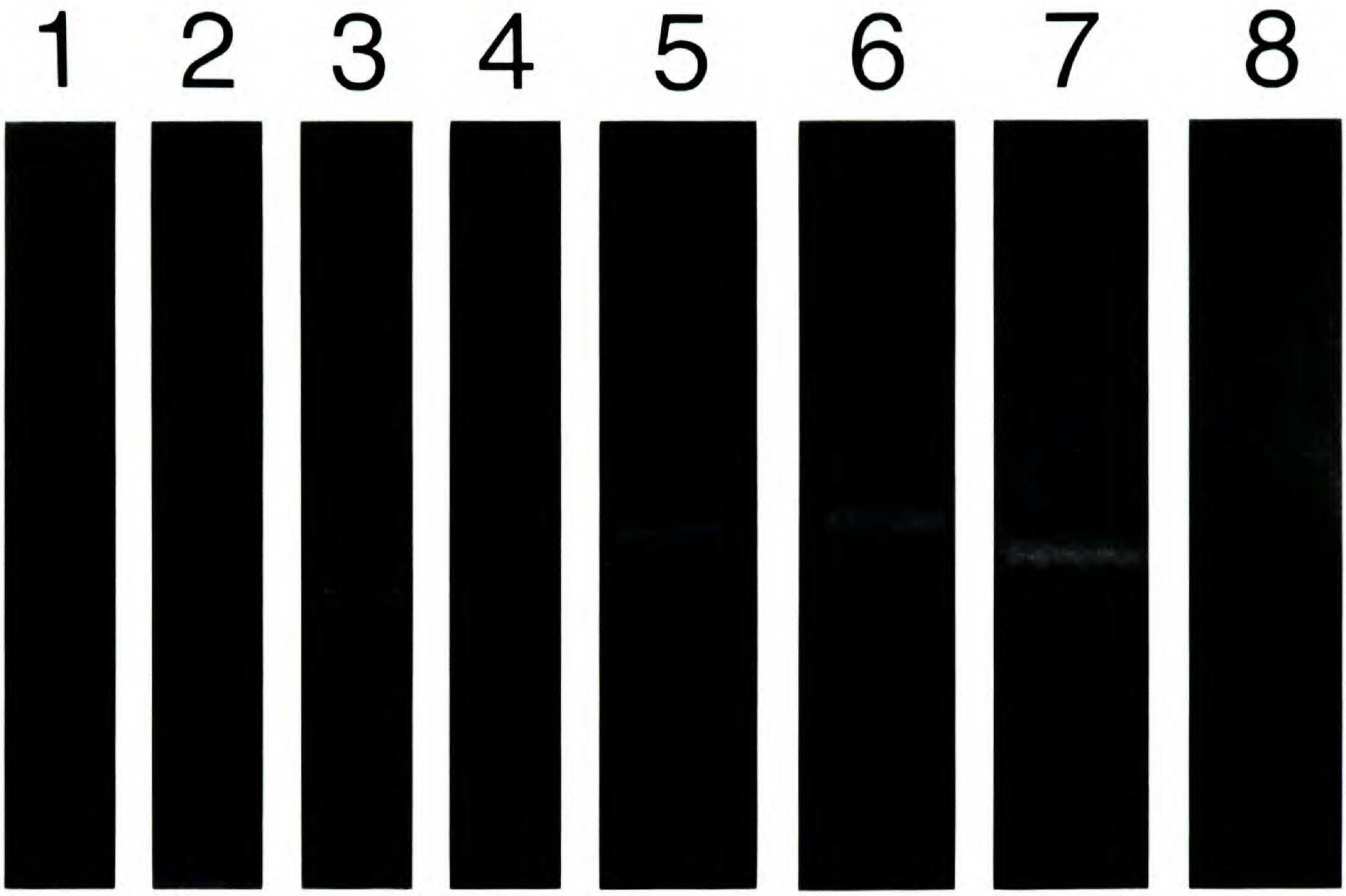

FIGURE 4. Agarose gel stained with ethidium bromide. PCR products amplified using primers for the gene encoding the chloroplast protein. Lane 1, *Brassica nigra*; 2, *B. napus*; 3, *R. schlechteri*; 4, *B. carinata*; 5, *B. oleracea*; 6, *B. juncea*; 7, *A. thaliana*; 8, *C. keyserri*. The *A. thaliana* PCR product is 590 bp.

through both restriction site and sequence analysis, due to its low rate of evolution. However, because chloroplasts are typically uniparentally inherited, the usefulness of cpDNA data sets may be limited when there has been hybridization and introgression. For example, some individuals of *Helianthus annuus* have the rDNA of one species and the cpDNA of another (Dorado et al., 1992; Rieseberg et al., 1990; Rieseberg & Brunsfeld, 1992). Phylogenetic relationships inferred from a tree based on cpDNA can contradict those inferred from an rDNA tree. Patterns of cpDNA distribution can also reflect geographical relationships rather than species designations (Brunsfeld et al., 1992; Matos, 1992; Soltis et al., 1992). Studies such as these make clear the need for additional nuclear markers for use in phylogenetic studies.

Many researchers have avoided using nuclear markers because of the concern that many nuclear genes are members of multi-gene families and that gene conversion among these genes would obscure orthologous (due to speciation and reflecting organismal phylogeny) and paralogous (due to gene duplication and reflecting gene phylogeny) relationships. In a recent review, Doyle (1993) stated that "The nuclear genome would appear to have limitless potential for phylogenetic studies . . . yet it remains largely untapped as a source of DNA characters." Gene duplication and conversion are widely believed to be important in the evolution of multi-gene families (Ohta, 1987, 1988). Frequent gene conversion among members of multi-gene families maintains sequence similarity but also restricts the independent evolution of individual genes and results in concerted evolution. Two forces will maintain high levels of similarity among gene sequences: selection and gene conversion. In gene conversion, related gene sequences are homogenized within a genome. Using computer simulation, Sanderson & Doyle (1992) examined the difficulty in determining true species relationships using sequence data from multi-gene families when there has been gene conversion and recombination. The results of this study suggest that the true species tree can be inferred when gene conversion is high (i.e., as in rDNA) and when gene conversion is low, but that intermediate levels of gene conversion will make it difficult to distinguish orthologous and paralogous genes and will obscure species relationships. An analysis of sequences of the nuclear gene

TABLE 3. Similarity of the nucleotide sequences of the genes for the chloroplast-localized proteins.

	P. sativum	*G. max*	*P. hybrida*	*A. thaliana*	*T. aestivum A*
P. sativum					
G. max	82.4%				
P. hybrida	65.7%	72.3%			
A. thaliana	63.1%	69.2%	63.3%		
T. aestivum A	52.2%	59.6%	54.7%	56.2%	
T. aestivum B	52.6%	59.2%	54.1%	55.5%	96.5%

family encoding the small subunit of ribulose 1,5-bisphosphatase carboxylase (*rbc*S) in 17 genera, including green algae and cyanobacteria, demonstrated gene conversion among the *rbc*S genes (Meagher et al., 1989). However, Meagher's analysis also indicates that in spite of gene conversion, *rbc*S sequences can be used in some instances to infer generic and higher-level relationships.

There is no evidence for widespread gene conversion across gene families of the small heat shock genes, although they appear to be related to each other through gene duplications. The small heat shock genes have a higher level of similarity within families across species than they do within species across families; this most likely reflects selective constraints. Similar results were reported in an analysis of the amino acid sequences of the small heat shock proteins (Vierling, 1991). Regions of high similarity within a genome are necessary for gene conversion to occur. Gene conversion would result in higher levels of similarity among genes across families, within a species. The small heat shock genes are evolving at a rate comparable to other nuclear genes and faster than genes encoded in the chloroplast (Wolfe et al., 1989). A comparison of the nuclear *Adh* genes from *Zea mays*, *Triticum aestivum*, and *Hordeum vulgare* found a Ks of 0.66 and a Ka of 0.03 (Wolfe et al., 1989). Between the Solanaceae and the Brassicaceae, the Ks was greater than 2.50, and Ka was 0.10 (Wolfe et al., 1989).

In a tree of representatives of all the small heat shock gene families, each gene family formed a well-supported clade (bootstrap values of 99–100%). However, there is a lack of support for some of the branches within the Class II cytosolic gene clade. In *Pharbatis nil* the 17.2 kDa protein is induced by changes in photoperiod and heat shock, and the 18.8 kDa protein is induced only by heat shock. Due to the lack of sequence data for the small heat shock genes in related species, it is unclear whether these two genes are the products of a recent or more ancient duplication. Hence, it is not possible to determine the paralogous and orthologous relationships between these genes and the other Class II sequences. It is this type of uncertainty that can make inferring organismal relationships problematical and perhaps unreliable.

The chloroplast-localized small heat shock genes may prove the most useful in phylogenetic analysis. These genes are present in a single copy in the Brassicaceae, and the gene tree reflects species relationships. In addition, these genes are longer than the other small heat shock genes and thus provide more characters. The sequences for the transit peptides were not used in these analyses due to the difficulty of aligning the sequences across distantly related taxa. However, the rapid evolution

TABLE 4. Rates of synonymous (Ks) and nonsynonymous (Ka) substitutions among the small heat shock genes.

	Ks	Ka
Z. mays vs. *T. aestivum*		
Chloroplast-localized	0.44 (± 0.13)	0.09 (± 0.02)
Class II Gene	0.36 (± 0.09)	0.08 (± 0.02)
A. thaliana vs. *P. hybrida*		
Chloroplast-localized	>2	0.16 (± 0.02)
A. thaliana vs. *G. max*		
Class I cytosolic	>2	0.21 (± 0.02)
Class II cytosolic	>2	0.22 (± 0.02)

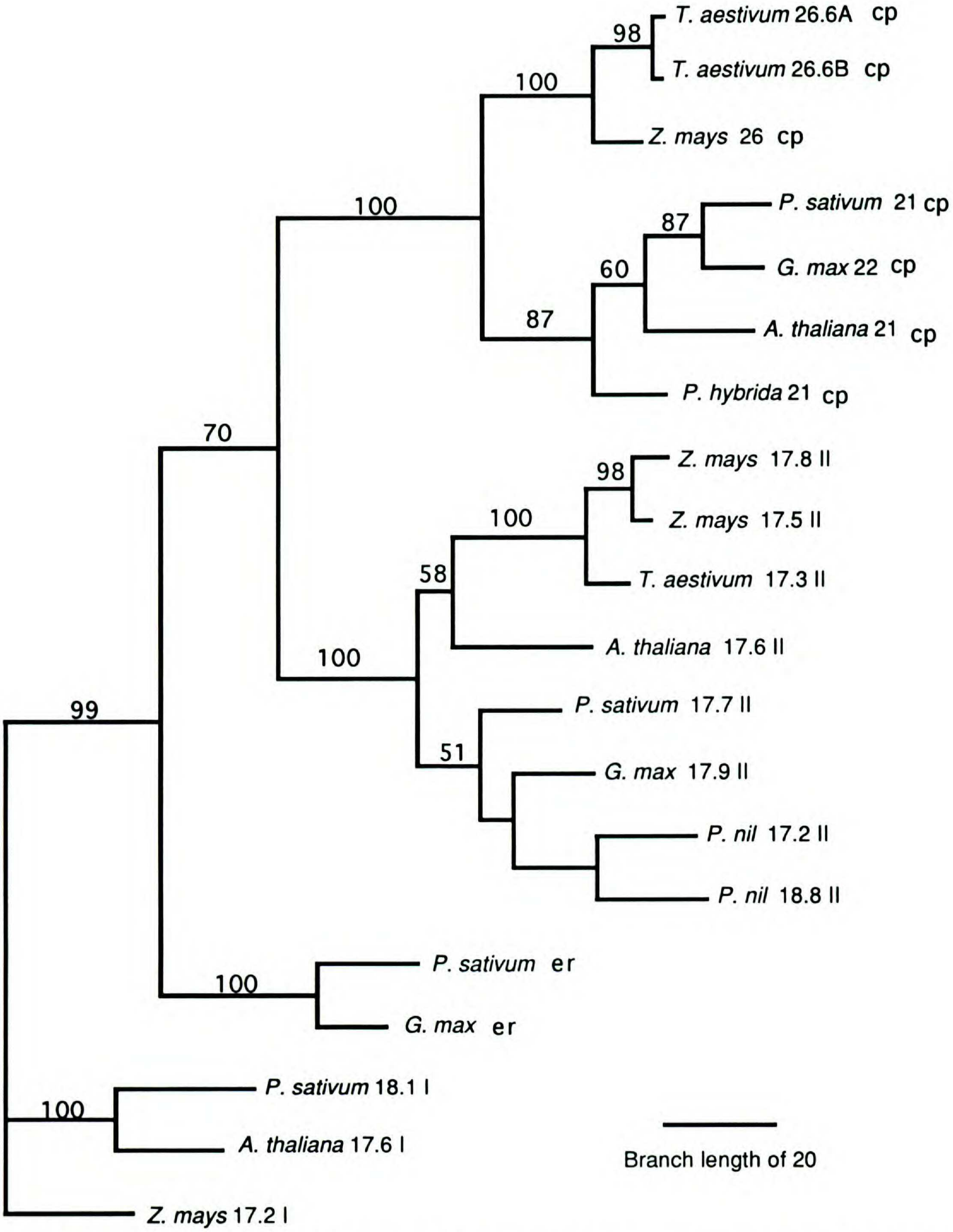

FIGURE 5. Tree of the small heat shock genes. The numbers above the branches are the numbers of times out of 100 bootstrap replicates that the branch was present. The names of the genes encoding the Class I cystosolic proteins are followed by I, those encoding the Class II cytosolic proteins by II, those encoding the endomembrane-localized proteins by er, and those encoding the chloroplast-localized proteins by cp.

of these sequences may prove informative in close comparisons.

For a gene to be useful in phylogenetic analysis the following four criteria need to be met. First, the gene should be present in a single copy within a genome. Second, there should be no variation below the species level but sufficient genetic variation between the species or taxa of interest. Third, if the gene is a member of a multi-gene family, orthologous and paralogous genes should be easily distinguishable. Fourth, the gene should have enough sequence similarity across the taxa of interest for Southern blot hybridization and primer annealing to be possible and no or few introns (Friendlander et al., 1992). In this paper, I have presented evidence that the small heat shock genes may be useful for phylogenetic analysis of plants. Restriction Fragment Length Polymorphism (RFLP) analysis indicates that the small heat shock genes analyzed in the Brassicaceae are single copy and are not variable below the species level but are variable between genera. The gene sequences of the four families of genes are readily distinguishable based on overall similarity comparisons and phylogenetic analysis. The sequences of the different gene families are sufficiently divergent (approximately 50%) that cross-hybridization in Southern blot analysis is unlikely, while within-family sequence conservation permits amplification of genes using PCR. However, not all the small heat shock genes may be useful. The orthologous and paralogous relationships among the genes for the Class II proteins may be complex. More sequences will be needed before this gene family can be fully evaluated for its utility in phylogenetic analysis.

DNA sequences for protein-coding nuclear genes may be very useful in inferring phylogenetic relationships, but they should not be used without careful examination of their evolutionary dynamics, including gene duplication and conversion. Of the four families of the small heat shock gene families, the genes coding for the chloroplast-localized protein are the most promising for phylogenetic study: a more detailed study of their usefulness in phylogenetic analysis is now in progress.

Literature Cited

Barling, D., H. Bulter, K. Liebeton & E. W. Weiler. 1992. An *Arabidopsis thaliana* cDNA clone encoding a 17.6 kDa Class II heat shock protein. Pl. Molec. Biol. 18: 1007–1008.

Becker, J. & E. A. Craig. 1994. Heat-shock proteins as molecular chaperones. Eur. J. Biochem. 219: 11–23.

Bousquet, J., S. H. Stauss, A. H. Doerksen & R. A. Price. 1992. Extensive variation in evolutionary rate of *rbc*L gene sequences. Proc. Natl. Acad. Sci. U.S.A. 89: 7844–7848.

Brunsfeld, S. J., D. E. Soltis & P. S. Soltis. 1992. Evolutionary patterns and processes in *Salix* sect. *Longifoliae*: Evidence from chloroplast DNA. Syst. Bot. 17: 239–256.

Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, Y.-L. Qiu, K. A. Kron, J. H. Rettig, E. Conti, J. D. Palmer, J. R. Manhart, K. J. Systma, H. J. Michaels, W. J. Kress, K. G. Karol, W. D. Clark, M. Hedrén, B. S. Gaut, R. K. Jansen, K.-J. Kim, C. F. Wimpee, J. F. Smith, G. R. Furnier, S. H. Strauss, Q.-Y. Xiang, G. M. Plunkett, P. S. Soltis, S. M. Swensen, S. E. Williams, P. A. Gadek, C. J. Quinn, L. E. Eguiarte, E. Golenberg, G. H. Learn, Jr., S. W. Graham, S. C. H. Barrett, S. Dayanandan & V. A. Albert. 1993. Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbc*L. Ann. Missouri Bot. Gard. 80: 528–580.

Chen, Q. & E. Vierling. 1991. Analysis of conserved domains identifies a unique structural feature of a chloroplast heat shock protein. Molec. Gen. Genet. 226: 425–431.

Clegg, M. T. & G. Zurawski. 1992. Chloroplast DNA and the study of plant phylogeny: Present status and future prospects. Pp. 1–13 *in* P. S. Soltis, D. E. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants. Chapman and Hall, New York.

Dixon, M. T. & D. M. Hillis. 1993. Ribosomal RNA secondary structure: Compensatory mutations and implications for phylogenetic analysis. Molec. Biol. Evol. 10: 256–267.

Dorado, O., L. H. Rieseberg & D. M. Arias. 1992. Chloroplast DNA introgression in southern California sunflowers. Evolution 46: 566–572.

Doyle, J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. Syst. Bot. 17: 144–163.

———. 1993. DNA, phylogeny, and the flowering of plant systematics. BioScience 43: 380–389.

Erickson, L. R., N. A. Straus & W. D. Beversdorf. 1983. Restriction patterns reveal origins of chloroplast genomes in *Brassica* amphidiploids. Theor. Appl. Genet. 65: 201–206.

Friendlander, T. P., J. C. Regier & C. Mitter. 1992. Nuclear gene sequences for higher level phylogenetic analysis: 14 promising candidates. Syst. Biol. 41: 483–490.

Gaut, B. S., S. V. Muse, W. D. Clark & M. T. Clegg. 1992. Relative rates of nucleotide substitutions at the *rbc*L locus of monocotyledonous plants. J. Molec. Evol. 35: 292–303.

Gething, M.-J. & J. Sambrook. 1992. Protein folding in the cell. Nature 355: 33–45.

Goping, I. S., J. R. H. Frappier, D. B. Walden & B. G. Atkinson. 1991. Sequence, identification and characterization of cDNAs encoding two different members of the 18 kDa heat shock family of *Zea mays* L. Pl. Molec. Biol. 16: 699–711.

Hamby, R. K. & E. A. Zimmer. 1992. Ribosomal RNA as a phylogenetic tool in plant systematics. Pp. 50–91 *in* P. S. Soltis, D. E. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants. Chapman and Hall, New York.

Helm, K. W. & E. Vierling. 1989. An *Arabidopsis*

thaliana cDNA clone encoding a low molecular weight heat shock protein. Nucl. Acids Res. 17: 7995.

———, P. R. LaFayette, R. T. Nagao, J. L. Key & E. Vierling. 1993. Localization of small heat shock proteins to the higher plant endomembrane system. Molec. Cell. Biol. 13: 238–247.

King, L. M. & B. A. Schaal. 1990. Genotypic variation within asexual lineages of *Taraxacum officinale*. Proc. Natl. Acad. Sci. U.S.A. 87: 998–1002.

Krishna, P., R. F. Felsheim, J. C. Larkin & A. Das. 1992. Structure and light-induced expression of a small heat-shock protein gene of *Pharbatis* Nil. Pl. Physiol. 100: 1772–1779.

Lauzon, L. M., K. Helm & E. Vierling. 1990. cDNA clone from *Pisum sativum* encoding a low molecular weight heat shock protein. Nucl. Acids Res. 18: 4274.

Li, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Molec. Evol. 36: 96–99.

Lindquist, S. & E. A. Craig. 1988. The heat-shock proteins. Annual Rev. Genet. 22: 631–677.

Matos, J. 1992. Evolution within the *Pinus montezumae* complex of Mexico: Population subdivision, hybridization and taxonomy. Ph.D. Dissertation. Washington University, St. Louis, Missouri.

Meagher, R. B., S. Berry-Lowe & K. Rice. 1989. Molecular evolution of the small subunit of ribulose bisphosphate carboxylase: Nucleotide substitution and gene conversion. Genetics 123: 845–863.

Ohta, T. 1987. Simulating evolution by gene duplication. Genetics 115: 207–213.

———. 1988. Multigene and supergene families. Pp. 41–65 *in* P. H. Harvey & L. Partridge (editors), Oxford Survey in Evolutionary Biology, Vol. 5. Oxford Univ. Press, New York.

Osteryoung, K., H. Sundberg & E. Vierling. 1993. Poly(A) tail length of a heat shock protein RNA is increased by severe heat stress, but intron splicing is unaffected. Molec. Gen. Genet. 239: 323–333.

Pakrash, S. & K. Hinata. 1980. Taxonomy, cytogenetics and origin of crop *Brassicas*, a review. Opera Bot. 55: 4–57.

Palmer, J. D., C. R. Shields, D. B. Cohen & T. J. Orton. 1983. Chloroplast DNA evolution and the origin of amphidiploid *Brassica* species. Theor. Appl. Genet. 65: 181–189.

Pamilo, P. & M. Nei. 1988. Relationships between gene trees and species trees. Molec. Biol. Evol. 5: 568–583.

Raaschke, E., G. Baumann & F. Schoffl. 1988. Nucleotide sequence analysis of soybean small heat shock protein genes belonging to two different multigene families. J. Molec. Biol. 199: 549–557.

Rieseberg, L. H. & S. J. Brunsfeld. 1992. Molecular evidence and plant introgression. Pp. 151–176 *in* P. S. Soltis, D. E. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants. Chapman and Hall, New York.

——— & D. E. Soltis. 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. Evol. Trends Pl. 5: 65–84.

———, S. Beckstrom-Sternberg & K. Doan. 1990. *Helianthus annuus* ssp. *texanus* has chloroplast DNA and nuclear ribosomal RNA genes of *Helianthus debilis* ssp. *cucmerifolius*. Proc. Natl. Acad. Sci. U.S.A. 87: 593–597.

Sambrook, J., E. F. Fritsch & T. Maniatis. 1989. Molecular Cloning: A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Sanderson, M. J. & J. J. Doyle. 1992. Reconstruction of organismal and gene phylogenies from data on multigene families: Concerted evolution, homoplasy, and confidence. Syst. Biol. 41: 4–17.

Schoffl, F. & J. L. Key. 1982. An analysis of mRNAs for a group of heat shock proteins of soybean using cloned cDNAs. J. Molec. Appl. Genet. 1: 301–314.

Smith, A. B. 1989. RNA sequence data in phylogenetic reconstruction: Testing the limits of its resolution. Cladistics 5: 321–344.

Soltis, D. E., P. S. Soltis & B. G. Milligan. 1992. Intraspecific chloroplast DNA variation: Systematic and phylogenetic implications. Pp. 117–150 *in* P. S. Soltis, D. E. Soltis & J. J. Doyle (editors), Molecular Systematics of Plants. Chapman and Hall, New York.

Susek, R. E. & S. L. Lindquist. 1989. hsp26 of *Saccharomyces cerevisiae* is related to the superfamily of small heat shock proteins but is without a demonstrable function. Molec. Cell. Biol. 9: 5265–5271.

Swofford, D. L. 1993. PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1. Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois.

Vierling, E. 1991. The heat shock response in plants. Annual Rev. Pl. Physiol. Pl. Molec. Biol. 42: 579–620.

———, R. T. Nagao, A. E. DeRocher & L. M. Harris. 1988. A heat shock protein localized to chloroplasts is a member of a eukaryotic superfamily of heat shock proteins. EMBO 7: 575–581.

Weng, J., Z. F. Wang & H. T. Nguyen. 1991. A *Triticum aestivum* cDNA clone encoding a low molecular weight heat shock protein. Pl. Molec. Biol. 16: 699–711.

Wolfe, K. H., P. M. Sharp & W.-H. Li. 1989. Rates of synonymous substitution in plant nuclear genes. J. Molec. Evol. 29: 208–211.

```
                          1                                                  50
T. aestivum 26.6A         CCTCGTCGCG CT---GGATT TGACATCTCC CCG------T TCGGGCTAGT
T. aestivum 26.6b         CCTCGTCGCG CT---GGATT TGACATCTCC CCG------T TCGGGCTAGT
P. sativum 21             CCTCGCAGAA GT---TCCAT CGATATTTCA CCA------T TTGGTTTATT
G. max 21                 CCACGTAGAA CT---GCCAT GGACATTTCA CCA------T TTGGTATCTT
A. thaliana 21            CAACAACGCT TA---ACCAT GGACGTCTCT CCT------T TCGGATTGTT
P. hybrida 21             CCTAGGAGAA TG---GCTCT TGATGTCTCC CCT------T TCGGGTTATT
Z. mays 17.8              GACGCGGTGA TGTTCGGGCT GGAGACT--- ---------- --------CC
Z. mays 17.5              GACGGGAGGA TGTTCGGTCT CGAGACC--- ---------- --------CC
T. aestivum 17.3          GCGGGCATGG TGTTCGGCTT GGATGCC--- ---------- --------CC
P. sativum 17.7           GATTTCAGGC TAATGGATTT GGATTCT--- ---------- --------CC
G. max 17.9               GATTTCAGAG TGATGGGTTT GGAGTCT--- ---------- --------CC
P. nil 17.2               GATTTGAGGT TGATGGGTTT TGATCAC--- ---------- --------CC
P. nil 18.8               GATCTCAGGA ACTTTGGCTT ATCCAACTTC GGGTTAGAAC CG---CAGCT
A. thaliana 17.6II        GATTTAGGAA GGTTT----- ---------- ---------- ----------
P. sativum 18.8           TCTCTGATTC CAAGT----- -TTCTTTAGT ---GGCCGAA GG------AG
A. thaliana 17.6          TCTCTAATTC CAAGC----- -ATCTTCGGA ---GGAAGAA GA------AC
Z. mays 17.2              TCGCTCGTG- ---------- ---------- ------AGGC GC------AG
P. sativum er             GCTGATTTTC CTTTGAAAGC AAAAGGATCA CTACTACCAT TCATAGAT--
G. max er                 ---------- -----AAAGC AAACGGGTCT CTGCTCCCAT TCATGGAT--

                                  51                                         100
T. aestivum 26.6A         GGACCCGATG TCGCCGATGA GGACGATGCG GCAGATGCTG GACACGATGG
T. aestivum 26.6B         GGACCCGATG TCGCCGATGA GGACGATGCG GCAGATGCTG GACACGATGG
P. sativum 21             GGACCCATGG TCACCTATGA GAAGCATGCG TCAGATGCTA GACACAATGG
G. max 21                 GGACCCGTGG TCACCCATGA GGAGCATGCG CCAGATATTG GACACCATGG
A. thaliana 21            GGATCCTTTG TCACCAATGA GGACGATGCG ACAAATGTTA GATACTATGG
P. hybrida 21             AGACCCAATG TCTCCCATGA GAACAATGAG GCAGATGATG GACACTATGG
Z. mays 17.8              CCTGATGGCG GCGCTGCAGC ACCTGCTGGA CGTGCCCGAC GGCGACGCCG
Z. mays 17.5              CCTGATGGTG GCGCTGCAGC ACCTTCTGGA CGTGCCCGAC GGCGACGCCG
T. aestivum 17.3          GATGATGGCC GCCCTGCAGC ACCTGCTGGA CATCCCGGAC GGCGAGGCGG
P. sativum 17.7           ACTCTTCAAC ACTCTCCATC ATATAATGGA CCTCACCGAC GAC---ACAA
G. max 17.9               ACTGTTCCAC ACGCTGCAAC ACATGATGGA CATGTCAGAG GAC---GGTG
P. nil 17.2               GCTCTTC--- ------CACC ACATCATGGA CTACGCCGGC GAC---GACA
P. nil 18.8               CCTCTCAACC ATCCAAGACA TGCTCGACTT CGCCGACGAC CACGACAGAG
A. thaliana 17.6II        -CCAATAATC TCAATCCTCG AAGACATGCT TGAAGTCCCC GAAGACCACA
P. sativum 18.8           CAATGTTTTC GATCCTTTCT CCCTGGAC-- -GTCTGGGAT CCTTTGAAGG
A. thaliana 17.6          AAACGTTTTC GATCCGTTCT CGCTGGAT-- -GTTTTCGAT CCGTTCGAAG
```

APPENDIX (pp. 289–295). DNA alignment of sequences listed in Table 1. The sequences coding for the transit peptides and the start ATG codons were removed and were not used in the analysis. A "-" denotes a gap in the alignment.

```
Z. mays 17.2          CAACGTGTTC GACCCCTTCT CGATGGAC-- -CTCTGGGAT CCCTTCGACA
P. sativum er         -------TCT CCCAACACTC TCTTATCGGA TCTCTGGTCT GATCGTTTCC
G. max er             -------CCT CCCATCACTC TCTTGGCTGA TCTCTGGTCC GATCGCTTCC

                      101                                                 150
T. aestivum 26.6A     ACCGGCTGTT CGACGACGCC GTGGGGTTCC CC------AC GCGTCGCTCG
T. aestivum 26.6B     ACCGGCTGTT CGACGACGCC GTGGGGTTCC CC------AC GGCGCGCTCG
P. sativum 21         ACAGAATTTT CGAAGATGCA ATCACAATTC CT---GGA-- -AGAAACATT
G. max 21             ACCGAGTTTT CGAAGATACC ATGACATTCC CT---GGA-- -AGAAACATA
A. thaliana 21        ACAGGATGTT CGAGGACACT ATGCCTGTCT CA---GGA-- -AGAAACAGA
P. hybrida 21         ACAGGCTATT CGAGGATACC ATGACATTCC CT---GGATC AAGGAACAGA
Z. mays 17.8          GCGCGGGTGG CGACAACAAG ACGGGCAGCG GCGGCAGCGC CACGCGCACC
Z. mays 17.5          GCGCGGGCGG CGACAAGGCG ------GGCG GCGGCGGGCC CACGCGCACC
T. aestivum 17.3      AGCCGCCGCC GGAGAAG--- ---------- --CAGGGCCC GACGCGCGCC
P. sativum 17.7       CCGAGAAGAA C--------- ---------T TAAACGCTCC AACTCGAACA
G. max 17.9           CAGGAGACAA CAAG------ ------ACAC ACAATGCTCC AACATGGTCA
P. nil 17.2           AGTCATCCAA CAGC------ ---------- --AGCGCGCC TTCTCGGACG
P. nil 18.8           CCGGCCGGGC C--------- ------CCAC CAGAACAACC CATCCGGGCC
A. thaliana 17.6II    ACAACGAGAA G--------- ------ACCC GCAACAACCC TTCACGAGTT
P. sativum 18.8       ACTTTCCATT TTCAAATTCT TCACCTTCCG CT---TCATT CCCT------
A. thaliana 17.6      GTTTCTTGAC GCCG---TCA GGATTGGCAA ACGCACCC-- -GCT------
Z. mays 17.2          CCATG---TT CCGCTCCATC GTCCCGTCGG CG-------- -ACCTCCACC
P. sativum er         CAGATCCGTT TCGCGTCTTA GAACAAATT- -----CCCTA TGGAGTTGAG
G. max er             CAGACCCGTT TCGCGTGCTG GAACATATT- -----CCGTT TGGGGTTGAC

                      151                                                 200
T. aestivum 26.6A     CCAGCGGCGC GAGCGAGA-- ----CGCCGG ATGCCGTGGG ACATCATGGA
T. aestivum 26.6B     CCGGCGCGGC GAGCGAAGAC GCCG---CGG ATGCCGTGGG ACATTATGGA
P. sativum 21         GGAGGA---- -----GGTGA GATT---CGT GTGCCTTGGG AAATCAAAGA
G. max 21             GGAGGA---- -----GGGGA GATC---CGT GCCCCTTGGG ACATCAAAGA
A. thaliana 21        GGAGGAAGTG GAGTGTCAGA GATT---CGT GCACCGTGGG ACATCAAAGA
P. hybrida 21         GGAACA---- -----GGGGA AATA---CGT GCCCCTTGGG ACATCAAAGA
Z. mays 17.8          TACGTCCGCG ACGCGCGCGC CATGGCGGCC ACCCCGGCCG ACGTGAAGGA
Z. mays 17.5          TACGTCGCCG ACGCGCGCGC CATGGCGGTC ACCCCGGCCG ACGTCAAGGA
T. aestivum 17.3      TACGTCCGCG ACGCGCGCGC CATGGCGGCC ACCCCCGCCG ACGTGAAGGA
P. sativum 17.7       TATGTCCGTG ACGCAAAGGC AATGGCTGCA ACTCCAGCGG ACGTGAAAGA
G. max 17.9           TACGTTCGAG ACGCGAAAGC AATGGCTGCA ACACCTGCGG ATGTGAAGGA
P. nil 17.2           TTTATGCTGG ACGCCAAGGC AATGGCAGCC ACCCCGGCGG ACGTGAAGGA
P. nil 18.8           TACGTCCGGG ACGCGAAGGC GATGGCAGCG ACCCCGGCGG ACGTGAAGGA
```

APPENDIX. Continued.

```
A. thaliana 17.6II   TACATGCGAG ACGCTAAGGC AATGGCTGCT ACACCTGCTG ACGTCATCGA
P. sativum 18.8      ------CGTG AGAATCCTGC TTTTGTGAGC ACACGAGTTG ACTGGAAGGA
A. thaliana 17.6     ------ATGG ATGTGGCAGC GTTCACAAAC GCTAAAGTGG ATTGGAGGGA
Z. mays 17.2         AAC---TCCG AGACTGCCGC CTTCGCCAGC GCCCGCATCG ACTGGAAGGA
P. sativum er        AAACACGAAC CATCCATAAC ATTGTCACAT GCTAGAGTAG ACTGGAAGGA
G. max er            AAAGATGAAG CATCCATGGC CATGTCACCT GCTAGAGTGG ACTGGAAGGA

                  201                                                  250
T. aestivum 26.6A    GGACGAGAAG GAGGTGAAGA TGCGGTTTGA CATGCCTGGG CTGTCGCGGG
T. aestivum 26.6B    GGACGAGAAG GAGGTGAAGA TGCGGTTTGA CATGCCTGGG CTGTCGCGGG
P. sativum 21        TGAAGAACAT GAAATCAGAA TGCGTTTTGA TATGCCTGGT GTTTCTAAGG
G. max 21            TGAAGAACAT GAAATCAGAA TGAGGTTTGA CATGCCGGGT CTTGCCAAGG
A. thaliana 21       GGAAGAACAC GAGATCAAGA TGCGTTTCGA CATGCCTGGT CTCTCTAAAG
P. hybrida 21        CGATGAAAAT GAAATCAAGA TGCGTTTTGA TATGCCAGGG CTATCTAAAG
Z. mays 17.8         GCTGCCCGGC GCGTACGCGT TCGTGGTGGA CATGCCGGGG CTGGGCACGG
Z. mays 17.5         GCTCCCGGGC GCGTACGCGT TCGTGGTGGA CATGCCGGGG CTGGGCACGG
T. aestivum 17.3     GCTGCCGGGC GCGTACGCGT TCGTGGTGGA CATGCCGGGG CTCGGGTCCG
P. sativum 17.7      GCATCCAAAT TCATACGTGT TTATGGTGGA CATGCCTGGG GTGAAATCTG
G. max 17.9          GTATCCGAAT TCTTACGTGT TCGAGATCGA CATGCCGGGT TTGAAATCTG
P. nil 17.2          GTACCCAAAC TCCTACGTCT TCATCATAGA CATGCCGGGG CTGAAGTCCG
P. nil 18.8          GTATCCGAAT TCCTACGTGT TCATCGCGGA CATGCCGGGG GTAAAGGCCG
A. thaliana 17.6II   GCACCCTAAC GCATATGCAT TCGTCGTGGA CATGCCTGGA ATCAAAGGAG
P. sativum 18.8      AACACCGGAA GCGCATGTTT TCAAGGCTGA TCTTCCTGGG CTGAAAAAGG
A. thaliana 17.6     GACACCTGAG GCGCACGTGT TCAAGGCGGA TTTACCGGGA CTGAGGAAGG
Z. mays 17.2         GACGCCCGAG GCGCACGTCT TCAAGGCCGA CCTCCCCGGC GTCAAGAAGG
P. sativum er        AACTCCAGAG GGACATGTGA TAATGGTGGA CGTGCCTGGG TTGAAAAAAG
G. max er            GACCCCAGAG GGGCATGTTA TAATGCTGGA CGTGCCGGGG CTGAAGAGAG

                     251                                                  300
T. aestivum 26.6A    AGGAGGTGAG GGTGATGGTG GAGGACGACG CGCTGGTCAT CCGCGGCGAG
T. aestivum 26.6B    AGGAGGTGAG GGTGATGGTG GAGGACGACG CGCTGGTCAT CCGCGGCGAG
P. sativum 21        AAGACGTTAA AGTATCAGTT GAAGATGATG TTCTTGTCAT AAAAAGTGAT
G. max 21            AAGATGTTAA GGTATCAGTG GAGGATGATA TGCTTGTTAT AAAAGGTGGC
A. thaliana 21       AAGACGTCAA AATCTCTGTA GAAGATAACG TACTTGTGAT CAAAGGAGAG
P. hybrida 21        AAGAGGTAAA AGTATCAGTG GAAGATGATG TGCTTGTCAT CAAAGGTGAA
Z. mays 17.8         GCGACATCCG GGTGCAGGTG GAGGACGAGC GGGTGCTGGT GGTCAGCGGC
Z. mays 17.5         GCGACATCAA GGTGCAGGTG GAGGACGAGC GGGTGCTGGT GATCAGCGGC
```

APPENDIX. Continued.

```
T. aestivum 17.3      GCGACATCAA GGTGCAGGTG GAGGACGAGC GGGTGCTGGT GATCAGCGGC
P. sativum 17.7       GTGACATAAA GGTTCAGGTG GAAGATGAGA ATGTGCTATT GATAAGTGGC
G. max 17.9           GGGACATAAA GGTTCAAGTT GAAGACGACA ACCTGCTTCT GATATGTGGG
P. nil 17.2           GCGACATCAA GGTGCAGGTC GACGGCGACA ACGTGTTGAG TATTAGCGGG
P. nil 18.8           CGGAAATCAA GGTCCAGGTG GAGGACGACA ACGTTTTGGT GGTGAGCGGG
A. thaliana 17.6II    ATGAGATCAA GGTTCAGGTC GAGAACGACA ATGTGCTTGT GGTGAGTGGA
P. sativum 18.8       AGGAAGTGAA AGTTGAAGTT GAAGATGATA GGGTTCTACA GATAAGCGGA
A. thaliana 17.6      AAGAGGTGAA AGTGGAGGTT GAGGATGGCA ACATACTTCA GATAAGTGGA
Z. mays 17.2          AGGAGGTCAA GGTTGAGGTC GAAGACGGCA ACGTGCTGGT CATCAGCGGC
P. sativum er         ATGATATAAA GATAGAAGTG GAAGAGAATA GGGTGCTAAG AGTGAGTGGT
G. max er             AAGAGATAAA GGTAGAGGTG GAAGAGAATA GGGTGCTGAG AGTGAGTGGT

                      301                                                350
T. aestivum 26.6A     CACAAGAAGG AG------GC CGGCGAAGGG CAGGGCGAAG GCGGCGACGG
T. aestivum 26.6B     CACAAGAAGG AG------GC CGGCGAAGGG CAGGGCGAAG GCGGCGACGG
P. sativum 21         CATAGA---- ---------- -GAGGAAAAT GGTGGAGAAG AT--------
G. max 21             CACAAAAGTG AA-------- -CAAGAACAT GGTGGAGATG AT--------
A. thaliana 21        CAGAAGAAGG AA-------- ---------- ---------G ACAGTGATGA
P. hybrida 21         CACAAAAAGG AA-------- ---------- ------AGTG GAAAGGATGA
Z. mays 17.8          GAGCGGCGCC GGGAG----- ----GAGCGC GAG---GACG ACGCCAAGTA
Z. mays 17.5          GAGCGGCGCC GGGAG----- ----GAGCGC GAG------G ACGCCAAGTA
T. aestivum 17.3      GAGCGGCGGA GGGAG----- ----GAGAAG GAG------G ACGCCAAGTA
P. sativum 17.7       GAGAGG---A AGAGA----- ----GAAGAA GAGAAAGAAG GTGTTAAATA
G. max 17.9           GAACGA---A AGAGG----- ----GACGAA GAGAAAGAAG GGGCGAAGTA
P. nil 17.2           GAGCGG---A AGCGT----- -GAGGCGGAG GAGAAGGAGG GGGCGAAGTA
P. nil 18.8           GAGCGAACGG AGAGG----- -GAGAAAGAC GAGAAGGATG GGGTGAAGTA
A. thaliana 17.6II    GAGAGGCAGA GAGAG----- -AACAAAGAA AAC---GAAG GTGTGAAGTA
P. sativum 18.8       GAGAGAAGCG TTGAG----- ----AAAGAA GAT---AAGA ATGATGAATG
A. thaliana 17.6      GAGAGGAGCA ATGAG----- ----AATGAA GAG---AAGA ATGACAAGTG
Z. mays 17.2          CAGCGCAGCA GGGAG----- ----AAGGAG GAC---AAGG ACGACAAGTG
P. sativum er         GAGAGGAAGA AAGAA----- ----GAAGAT AAA---AAAG GAGATCATTG
G. max er             GAAAGGAAGA AGGAG----- ----GAGGAG AAG---AAGG GGGATCACTG

                      351                                                400
T. aestivum 26.6A     GTGGTGGAAG GAGCGCAGCG TGAGCTCCTA CGACATGCGC CTTGCTCTG-
T. aestivum 26.6B     GTGGTGGAAG GAGCGCAGCT TGAGCTCCTA CGACATGCGA CTGGCTCTG-
P. sativum 21         -TGTTGGTCA AGAAAGAGTT ATAGTTGTTA TGATACTCGT CTCAAGCTT-
G. max 21             -TCTTGGTCA AGCAGGACCT ATAGTTCCTA TGATACCCGT TTAAAGCTG-
P. hybrida 21         CTCGTGGGGA AGG---AATT ACAGTTCCTA TGACACTCGT CTAAGTTTA-
```

APPENDIX. Continued.

```
Z. mays 17.8          CCTGCGCATG GAGCGGCGGA TGGGCAAGTT CATGCGCAAG TTCGTGCTG-
Z. mays 17.5          CCTGCGCATG GAGCGGCGGA TGGGCAAGTT CATGCGCAAG TTCGTGCTG-
T. aestivum 17.3      CCTGCGGATG GAGCGCCGCA TGGGCAAGCT GATGCGCAAG TTCGTGCTG-
P. sativum 17.7       TTTGAAGATG GAAAGAAGGA TTGGTAAGTT GATGAGGAAA TTTGTGTTA-
G. max 17.9           TTTGAGAATG GAGAGAAGGG TTGGGAAGTT AATGCGCAAG TTTGTGCTG-
P. nil 17.2           TGTGAGGATG GAGAGGAGGG TGGGGAAGTT AATGAGGAAG TTTGTGCTG-
P. nil 18.8           TTTGAGGATG GAACGGAGGG TGGGGAAGTT TATGAGGAAA TTTGTGCTG-
A. thaliana 17.6II    TGTGAGGATG GAGAGAAGAA TGGGTAAGTT CATGAGGAAG TTTCAGTTG-
P. sativum 18.8       GCATCGCGTG GAACGTAGCA GTGGAAAGTT CTTAAGAAGG TTCAGATTG-
A. thaliana 17.6      GCACCGTGTG GAGCGGTCGA GTGGGAAGTT CACGAGGAGG TTTAGGCTG-
Z. mays 17.2          GCACCGTGTC GAGCGCAGCA GTGGCCAGTT CATCAGGCGC TTCCGCCTG-
P. sativum er         GCACAGAGTT GAAAGATCTT ATGGAAAGTT CTGGAGGCAG TTTAAATTA-
G. max er             GCACAGAGTG GAGAGATCCT ATGGCAAGTT CTGGAGGCAG TTCAGGTTG-

                  401                                                 450
T. aestivum 26.6A     --CCGGACGA GTGCGACAAG AGCCAGGTGC GGGCCGAGCT CAAGAACGGC
T. aestivum 26.6B     --CCGGACGA GTGCGACAAG AGCCAGGTGC GCGCCGAGCT CAAGAACGGC
P. sativum 21         --CCTGATAA CTGTGAGAAA GAGAAGGTTA AGGCTGAGTT GAAAGATGGT
G. max 21             --CCGGATAA TTGCGAGAAG GACAAGGTTA AGGCAGAGTT GAAAAATGGT
A. thaliana 21        --CCAGACAA CTGTGAGAAA GACAAGATCA AAGCTGAGCT CAAGAACGGA
P. hybrida 21         --CCTGATAA TGTTGACAAA GATAAAGTTA AAGCAGAATT GAAGAATGGA
Z. mays 17.8          --CCGGACAA CGCCGACGTC GACAAGGTCG CCGCCGTGTG CAGGGACGGC
Z. mays 17.5          --CCGGACAA CGCCGACATG GACAAGATCT CGGCGGTGTG CAGGGACGGC
T. aestivum 17.3      --CCGGAGAA CGCCGACATG GAGAAGATCT CGCCG---TG CCGCGACGGC
P. sativum 17.7       --CCTGAGAA TGCGAATATT GAAGCTATCT CTGCTATTTC TCAAGATGGT
G. max 17.9           --CCTGAGAA TGCCAACACT GATGCAATCT CTGCTGTGTG CCAAGATGGT
P. nil 17.2           --CCGGAAAA TGCGAATAAG GAGAAGATCA CGGCGGTTTG CCAGGATGGG
P. nil 18.8           --CCGGAGAA TGCCAACGTG GAGGCGATAA ATGCGGTTTA TCAGGACGGG
A. thaliana 17.6II    --CCTGAAAA TGCAGATTTG GACAAGATCT CTGCTGTTTG TCATGACGGT
P. sativum 18.8       --CCTGAGAA TGCTAAAATG GATAAAGTGA AAGCTTCCAT GGAGAACGGC
A. thaliana 17.6      --CCGGAGAA TGCTAAGATG GAGGAAATCA AAGCCAGTAT GGAAAATGGT
Z. mays 17.2          --CCGGATGA CGCCAAGGTG GATCAGGTCA AGGCTGGCCT CGAGAACGGC
P. sativum er         --CCTCAAAA TGTTGATTTG GATTCTGTCA AAGCTAAAAT GGAAAACGGT
G. max er             --CCACAAAA TGTAGACTTG GATTCTGTCA AGGCCAAGCT GGAGAATGGG

                      451                                                 500
T. aestivum 26.6A     GTGCTGCTCG TGTCCGTGCC CAAGAGG--- ---------- ----------
T. aestivum 26.6B     GTGCTGCTCG TGTCCGTGCC CAAGAGG--- ---------- ----------
P. sativum 21         GTGCTTTATA TTACTATTCC TAAGACT--- ---------- ----------
```

APPENDIX. Continued.

```
G. max 21              GTGCTTTATA TCACCATTCC TAAGACC--- ---------- ----------
A. thaliana 21         GTCCTCTTTA TCACTATCCC TAAGACC--- ---------- ----------
P. hybrida 21          GTTCTTTTAA TCTCAATTCC TAAGACC--- ---------- ----------
Z. mays 17.8           GTGCTCACGG TGACCGTTGA GAAGCTGCCC ---------- ----------
Z. mays 17.5           GTGCTCACCG TGACCGTCGA GAAGCTGCCC ---------- ----------
T. aestivum 17.3       GTGCTCACCG TGACCGTCGA CAAGCTGCCG ---------- ----------
P. sativum 17.7        GTTCTTACGG TTACAGTTAA TAAATTGCCT ---------- ----------
G. max 17.9            GTGCTTAGTG TAACCGTGCA GAAATTGCCT ---------- ----------
P. nil 17.2            GTGTTGACGG TGACGGTGGA GAACGTGCCG ---------- ----------
P. nil 18.8            GTGTTGCAGG TGACGGTGGA GAAATTGCCG ---------- ----------
A. thaliana 17.6II     GTGTTGAAGG TTACTGTTCA GAAACTTCCT ---------- ----------
P. sativum 18.8        GTTCTGACAG TGACCGTTCC AAAA---GAA GAG------- ----------
A. thaliana 17.6       GTGTTGTCGG TTACCGTGCC CAAG---GTG CCA------- ----------
Z. mays 17.2           GTGCTCACGG TCACCGTGCC TAAG---GCG ---------- ----------
P. sativum er          GTTCTTACTT TAACTCTTCA TAAGTTGTCG CATGATAAGA TTAAAGGTCC
G. max er              GTGCTCACTT TGACACTTGA CAAGTTGTCA CCTGGTAAGA TCAAAGGTCC

                       501                                                 550
T. aestivum 26.6A      ---------- ---------- --------GA GACCGAGCGC AAGGTCATCG
T. aestivum 26.6B      ---------- ---------- --------GA GACCGAGCGC AAGGTCATCG
P. sativum 21          ---------- ---------- --------AA AATTGAACGC ACGGTTATTG
G. max 21              ---------- ---------- --------AA GGTTGAACGT AAGGTTATTG
A. thaliana 21         ---------- ---------- --------AA AGTCGAACGC AAAGTCATCG
P. hybrida 21          ---------- ---------- --------AA AGTTGAGAAG AAGGTGACTG
Z. mays 17.8           ---------- ---------- -----CCGCC GGAGCCCAAG AAGCCCAAGA
Z. mays 17.5           ---------- ---------- -----CCGCC CGAGCCCAAG AAGCCCAAGA
T. aestivum 17.3       ---------- ---------- -----CCGCC CGAGCCCAAG AAGCCCAAGA
P. sativum 17.7        ---------- ---------- -----CCACC TGAACCTAAG AAACCAAAAA
G. max 17.9            ---------- ---------- -----CCACC TGAGCCTAAG AAACCTAGGA
P. nil 17.2            ---------- ---------- -----CCGCC GGAACCGAAG AAGCCAAGGA
P. nil 18.8            ---------- ---------- -----CCGCC GGAGCCGAAG AAGCCAAAGA
A. thaliana 17.6II     ---------- ---------- -----CCTCC GGAACCAAAG AAACCAAAGA
P. sativum 18.8        ---------- ---------- -----ATAAA GAAGGCTGAG GTTAAGTCTA
A. thaliana 17.6       ---------- ---------- -----GAGAA AAAGCCAGAG GTTAAGTCCA
Z. mays 17.2           ---------- ---------- -----GAAGA GAAGAAGCCT GAGGTGAAGG
P. sativum er          TAGAATGGTT AGTATTGTGG AAGAGGATGA CAAACCATCT AAGATCGTCA
G. max er              CAGGGTGGTC AGCATTGCTG GGGAGGATCA CCAACAAGGT AATCTCAACA

                       551                 573
```

APPENDIX. Continued.

```
T. aestivum 26.6A    ACGTGCAGGT CCAG------ ---
T. aestivum 26.6B    ACGTGCAGGT CCAG------ ---
P. sativum 21        ATGTTCAAAT TCAG------ ---
G. max 21            ATGTCCAAGT TCAG------ ---
A. thaliana 21       ATGTCCAGAT TCAG------ ---
P. hybrida 21        ACGTGGAA-- ---------- ---
Z. mays 17.8         CCATCGAGGT CAAGGTCGCC ---
Z. mays 17.5         CCATCGAGGT CAAGGTCGCC ---
T. aestivum 17.3     CCATCCAGGT CCAGGTCGCC ---
P. sativum 17.7      CTATTCAAGT TAAGGTTGCT ---
G. max 17.9          CTATTCAGGT TAAGGTTGCT ---
P. nil 17.2          CCATTGAGGT CAAAATTGGT ---
P. nil 18.8          CAGTTGAAGT CAAGGTTGCG ---
A. thaliana 17.6II   CAATTCAAGT TCAAGTTGCT ---
P. sativum 18.8      TTGAGATTTC TGGT------ ---
A. thaliana 17.6     TTGATATCTC TGGT------ ---
Z. mays 17.2         CTATTGAGAT CTCTGGT--- ---
P. sativum er        ATGATGAGTT GAAA------ ---
G. max er            ATGATGGGGC CAAGCAGGAG CTT
```

APPENDIX. Continued.